Bayesian inference: an introduction
Some applications

Peter Green

School of Mathematics
University of Bristol

8/9 September 2011 / MLSS 2011, Bordeaux
Outline

6 Protein matching and alignment
   • Modelling
   • Inference
   • Results

7 Emission tomography
   • SPECT
   • Inverse problems
   • Geometrical perspective
   • Consistency and rates of convergence

8 Sparse latent factor analysis
   • Modelling
   • Computation
   • Model choice
   • Arabidopsis transcripts and metabolites
Protein matching and alignment

Joint work with K. V. Mardia (Leeds)


Matching and alignment: what we wish to do

A key example: aligning active sites of proteins.

Raw data: first two dimensions
Matching and alignment: what we wish to do

Two point configurations separately rotated to optimal alignment, with optimal matching.
Aligning protein gels
Aligning protein gels

Two point configurations (spots on two gels): can you see which points align with which?
Aligning protein gels

The 17 most probable matches in the gel data, the optimal match using our methodology – includes all 10 expert-identified matches
An important problem in shape analysis: aligning two or more configurations of points in space – inferring the geometrical transformation mapping one to the other.

Often, the configurations are not (fully) labelled, so that we have to match the configurations as well. An additional and interesting complication is when only (unknown) subsets of the configurations are to be matched.

Uncertainties about the two unknowns – the transformation and the subset matching – are inter-related. It is appealing, perhaps essential to infer these two unknowns simultaneously. We set up a probability model to do this.
Point process model, with geometrical transformation and random thinning

Given two point configurations in \( d \)-dimensional space \( \mathbb{R}^d \):
\[
\{x_j, j = 1, 2, \ldots, m\} \text{ and } \{y_k, k = 1, 2, \ldots, n\}
\]
— labelled for identification, but arbitrarily.

There may be a (possibly unknown) geometrical transformation \( A \) between the \( x \)-space and the \( y \)-space. Both sets regarded as noisy observations on hidden true locations \( \{\mu_i\} \): we do not know the mappings from \( j \) and \( k \) to \( i \).

Want to make model-based inference about these mappings, including probability statements about matching — which pairs \( (j, k) \) correspond to the same true location?
Point process model, with geometrical transformation and random thinning

Given two point configurations in $d$-dimensional space $\mathbb{R}^d$: 
\{x_j, j = 1, 2, \ldots, m\} and \{y_k, k = 1, 2, \ldots, n\} – labelled for identification, but arbitrarily.

There may be a (possibly unknown) geometrical transformation $A$ between the $x$-space and the $y$-space. Both sets regarded as noisy observations on hidden true locations $\{\mu_i\}$: we do not know the mappings from $j$ and $k$ to $i$.

Want to make model-based inference about these mappings, including probability statements about matching – which pairs $(j, k)$ correspond to the same true location?
Point process model, with geometrical transformation and random thinning

Given two point configurations in $d$-dimensional space $\mathcal{R}^d$: 
\{x_j, j = 1, 2, \ldots, m\} and \{y_k, k = 1, 2, \ldots, n\} – labelled for identification, but arbitrarily.

There may be a (possibly unknown) geometrical transformation $A$ between the $x$-space and the $y$-space. Both sets regarded as noisy observations on hidden true locations $\{\mu_i\}$: we do not know the mappings from $j$ and $k$ to $i$.

Want to make model-based inference about these mappings, including probability statements about matching – which pairs $(j, k)$ correspond to the same true location?
Identifying points

The mappings between the indexing of the \( \{ \mu_i \} \) and that of the data \( \{ x_j \} \) and \( \{ y_k \} \) are captured by indexing arrays \( \{ \xi_j \} \) and \( \{ \eta_k \} \); specifically we assume that

\[
x_j = \mu \xi_j + \varepsilon_1 \quad \text{and} \quad A y_k = \mu \eta_k + \varepsilon_2
\]

for \( j = 1, 2, \ldots, m \), \( k = 1, 2, \ldots, n \), where \( \{ \varepsilon_{rj} \} \) have probability density \( f_r \).

All \( \{ \varepsilon_1 \} \) and \( \{ \varepsilon_2 \} \) are independent of each other, and independent of the \( \{ \mu_i \} \).

‘\( x_j \) matches \( y_k \)’ means that \( \xi_j = \eta_k \).
The matching matrix

The matching of the configurations is represented by the matching matrix $M$: $M_{jk}$ indicates whether $x_j$ and $y_k$ are matched, i.e.

$$M_{jk} = \begin{cases} 1 & \text{if } \xi_j = \eta_k \\ 0 & \text{otherwise} \end{cases}.$$ 

Note that $\sum_{j,k} M_{jk}$ is the number of matches.

Each point can match at most one point in the other configuration, so there is at most one ‘1’ in each row and each column.
Suppose that

- the set of true locations \( \{ \mu_i \} \) forms a homogeneous Poisson process with rate \( \lambda \) over a region of volume \( v \)
- independently for each point, these may give rise to \( x \) or \( y \) points, both, or neither. Let \( \rho \) be the ratio \( P\{ \text{both} \} / P\{ \text{only } x \} P\{ \text{only } y \} \).
- all matching matrices with the same number of matches have the same probability

Then we can show

\[
p(M) \propto (\rho / \lambda v)^{\sum_{j,k} M_{jk}}
\]
Prior for $M$, based on Poisson process

Suppose that

- the set of true locations $\{\mu_i\}$ forms a homogeneous Poisson process with rate $\lambda$ over a region of volume $v$
- independently for each point, these may give rise to $x$ or $y$ points, both, or neither. Let $\rho$ be the ratio $P\{\text{both}\}/P\{\text{only } x\} P\{\text{only } y\}$.
- all matching matrices with the same number of matches have the same probability

Then we can show

$$p(M) \propto \left(\frac{\rho}{\lambda v}\right)^\Sigma j,k M_{jk}$$
Where does this come from?

Recall that: if you thin a Poisson process (of rate $\lambda$ in a region of length/area/volume $v$), retaining each point independently with probability $p$, then the number of retained points $\sim$ Poisson($\lambda vp$).

Here we have two dependent thinnings of the same process, and if we let $N_x$, $N_y$ and $N_{\text{both}}$ stand for the numbers of ‘$x$ only’, ‘$y$ only’ and ‘both $x$ and $y$’ points, we want

$$P\{N_{\text{both}} = z | N_x + N_{\text{both}} = m, N_y + N_{\text{both}} = n\}$$

where $N_x$, $N_y$, $N_{\text{both}}$ are independent Poisson distributed.

Exponentials cancel...
Likelihood of data

For simplicity, we now suppose $\mathcal{A}$ to be an affine transformation: $\mathcal{A}y = Ay + \tau$.

Making full use of the Poisson process assumption, let us integrate out the $\{\mu_i\}$: we find

$$p(x, y|M, A) = v^{-(m+n)}|A|^n \prod_{j,k: M_{jk}=1} g(x_j - Ay_k - \tau).$$

where $g(z) = \int f_1(z + u)f_2(u)du$ (the density of $\varepsilon_{1j} - \varepsilon_{2k}$).
Likelihood of data

For simplicity, we now suppose $\mathcal{A}$ to be an affine transformation: $\mathcal{A}y = Ay + \tau$.

Making full use of the Poisson process assumption, let us integrate out the $\{\mu_i\}$: we find

$$p(x, y|M, \mathcal{A}) = v^{-(m+n)}|A|^n \prod_{j,k: M_{jk}=1} g(x_j - Ay_k - \tau).$$

where $g(z) = \int f_1(z + u)f_2(u)du$ (the density of $\varepsilon_{1j} - \varepsilon_{2k}$).
Where does this come from?

Recall that: given that $N$ points of a Poisson process are realised in a region, the points are independently uniformly distributed over that region.

So we can marginalise out the hidden points $\mu_i$ by separate integrals for each...

Once we do so, the unmatched points ‘disappear’ – they carry no information about the parameters (or to put it another way, a Poisson process with points independently additively perturbed is still Poisson).
Spherical normal case

If we further specialise by making assumptions that all errors are $N(0, \sigma^2)$, then

$$g(z) = g_\sigma(z) = \frac{1}{(\sigma \sqrt{2})^d} \phi(z/\sigma \sqrt{2})$$

where $\phi$ is the standard normal density in $\mathcal{R}^d$. 
Joint distribution of data and unknowns

Our final joint model is

\[ p(M, A, \tau, \sigma, x, y) \propto |A|^n p(A) p(\tau) p(\sigma) \]

\[ \times \prod_{j,k:M_{jk}=1} \{ (\rho/\lambda) g_\sigma(x_j - Ay_k - \tau) \} \]

Each factor in the product expresses the competing explanations for the proximity of \( x_j \) and \( Ay_k + \tau \): matching \( (g_\sigma(\cdot)) \) and coincidentally close points \( (\lambda/\rho) \)…
Our final joint model is

\[ p(M, A, \tau, \sigma, x, y) \propto |A|^n p(A)p(\tau)p(\sigma) \times \prod_{j, k: M_{jk} = 1} \{(\rho/\lambda)g(\sigma)(x_j - Ay_k - \tau)\} \]

This will be the basis for our inference about \( M, A, \tau \) and \( \sigma \) – we have various options. Note that only the ratio \( \rho/\lambda \) is identifiable.
Inference under this model

EM – for either MLE or MAP – is an option; but it would allow us to study only certain aspects of an approximate version of our model, and is not trivial numerically (unique matching constraint, interplay of 7 real parameters).

Obtaining complete posterior by MCMC sampling gives a lot more freedom in inference.

Here we try only uninformative priors, some chosen for convenient updating.
Conjugate prior for a rotation

Rigid-body case: $A$ is a rotation, so $|A| = 1$, and it turns out that $p(A|M, \tau, \sigma, x, y)$

$$
\propto p(A) \exp \left( \text{tr} \left\{ (1/2\sigma^2) \sum_{j, k: M_{jk} = 1} y_k (x_j - \tau)^T A \right\} \right)
$$

So if $p(A)$ has the form $p(A) \propto \exp(\text{tr}(F_0^T A))$ for some matrix $F_0$, then so does the posterior, but $F_0$ becomes

$$
F = F_0 + (1/2\sigma^2) \sum_{j, k: M_{jk} = 1} (x_j - \tau)y_k^T
$$

This is known as the matrix Fisher distribution.
Updating for matching matrix $M$

$M$ is updated in detailed balance using Metropolis-Hastings moves that only propose simple changes:

(a) adding a match: changing one $M_{jk}$ from 0 to 1
(b) deleting a match: changing one $M_{jk}$ from 1 to 0
(c) switching a match: simultaneously changing one entry from 0 to 1, and another *in the same row or column* from 1 to 0.

For these, $p(M^{\text{prop}}, A, \tau, \sigma|x, y)/p(M, A, \tau, \sigma|x, y)$ has only a few factors, so updating is efficient.
Loss functions

The prior and likelihood determine the joint posterior of transformation $(A, \tau)$, matching ($M$) and variance parameter ($\sigma$) which we are free to summarise/visualise to taste.

If a point estimate of any unknowns is needed, we must specify a loss function. (We are uncomfortable about blindly reporting posterior mode or mean).

We concentrate here on the matching $M$, to be estimated by $\hat{M}$ minimising $E[L(M, \hat{M})|x, y]$. Other choices (additive over possible matches) are possible, but suppose we incur cost $\ell_{01}$ for each pair $(x_j, y_k)$ we falsely declare to be a match, and $\ell_{10}$ for each true match we fail to declare.
Loss functions

The prior and likelihood determine the joint posterior of transformation $(A, \tau)$, matching $(M)$ and variance parameter $(\sigma)$ which we are free to summarise/visualise to taste.

If a point estimate of any unknowns is needed, we must specify a loss function. (We are uncomfortable about blindly reporting posterior mode or mean).

We concentrate here on the matching $M$, to be estimated by $\hat{M}$ minimising $E[L(M, \hat{M})|x, y]$. Other choices (additive over possible matches) are possible, but suppose we incur cost $\ell_{01}$ for each pair $(x_j, y_k)$ we falsely declare to be a match, and $\ell_{10}$ for each true match we fail to declare.
Optimal matching

Providing $\ell_{10} + \ell_{01} > 0$ and $\ell_{01} > 0$, the optimal matching $\hat{M}$ is easily shown to be that maximising

$$\sum_{j,k: \hat{M}_{jk} = 1} (p_{jk} - K)$$

(penalised sum of marginal posterior match probabilities $p_{jk} = p(M_{jk} = 1|x, y)$) where $K = \ell_{01}/(\ell_{10} + \ell_{01})$.

This weighted bipartite matching problem is equivalent to a mathematical programming assignment problem, and can be solved by special-purpose or general LP methods (or for small problems, by inspection).
Optimal matching

Providing $\ell_{10} + \ell_{01} > 0$ and $\ell_{01} > 0$, the optimal matching $\hat{M}$ is easily shown to be that maximising

$$\sum_{j,k: \hat{M}_{jk} = 1} (p_{jk} - K)$$

(penalised sum of marginal posterior match probabilities $p_{jk} = p(M_{jk} = 1 | x, y)$) where $K = \ell_{01} / (\ell_{10} + \ell_{01})$.

This weighted bipartite matching problem is equivalent to a mathematical programming assignment problem, and can be solved by special-purpose or general LP methods (or for small problems, by inspection).
Matching protein gels

Wish to match two electrophoretic gels automatically, given point locations of centres of 35 proteins on each gel.
Matching protein gels

The 17 most probable matches in the gel data, the optimal match for any $K \in (0.3998, 0.7552)$ – includes all 10 expert-identified matches.
Aligning proteins in 3 dimensions
Aligning proteins in 3 dimensions

Why?

Matched ‘functional’ sites of two proteins suggest that they have similar functions - so if we do not know the function of one protein, we can infer it from that of the other. It then, e.g., helps in making a cheaper drug, assuming one protein is easier to manufacture chemically than the other.

Also we can infer evolutionary information.
Aligning proteins in 3 dimensions

Matching ‘active sites’ of two proteins – 17–beta hydroxysteroid dehydrogenase and carbonyl reductase.

Point configurations are centres of gravity of amino-acids: \( m = 40 \), \( n = 63 \).

From chemical properties of the sites, the relevant matching should be invariant under rigid transformation.
Aligning proteins

Raw data: first two dimensions
Aligning proteins

Raw data: dimensions 1 & 3
Aligning proteins

Raw data: dimensions 2 & 3
Aligning proteins

Two point configurations separately rotated to optimal alignment, with optimal matching when $K = 0.5$ (36 matches)
Mixing

![Graphs showing the results of protein matching and alignment](image)

- **# matches**
- **log sigma**
- **tau 1, tau 2, tau 3**
- **theta 12, theta 13, theta 23**
MCMC issues

- Multiple starts
- Conjugacy $\Rightarrow$ possibility of integrating out $A, \tau, \sigma$
  (matched-configuration-specific normalising constants)
- Tempering
Aligning proteins

Posterior probabilities of each match.
Note that we did not use sequence information!

‘Sequence order-independent structural comparison’, implications for evolution and protein folding.
Aligning proteins

Rotated to posterior mean transformation, showing sequence order and our matching.
Extension 1: Using partial labelling information

Sometimes points carry marks or ‘colours’ – quantitative information influencing tendency to match: handle by extending data model to generate colours as well as locations of $x$ and $y$ points, with colours (usually positively) dependent, independent of locations.

If we choose

$$\frac{P\{\text{obs x coloured } r \text{ and y coloured } s\}}{P\{\text{obs x coloured } r\}P\{\text{obs y coloured } s\}} = \rho \exp\{\gamma I[r = s] + \delta I[r \neq s]\}$$

then the extended model is fitted by simple multiplicative changes to Metropolis ratios in updates of $M$. 


Ext’n 2: Matching multiple configurations

Extend model to arbitrary numbers of configurations, inferring matching of any subsets.
Extension to more than 2 configurations

Ingredients:

- locations \( \{ \mu_i \} \),
- for \( c = 1, 2, \ldots, C \), configuration \( x^{(c)} \) with \( n_c \) occurrences,
- labelling vectors \( \xi^{(1)}, \ldots, \xi^{(C)} \),
- transformations \( A^{(1)}, \ldots, A^{(C)} \). Once again we have
  \[
  A^{(c)} x_j^{(c)} = A^{(c)} x_j^{(c)} + \tau^{(c)} \quad \text{for} \quad j = 1, \ldots, n_c, \quad c = 1, \ldots C.
  \]
Joint model for parameters and data

The joint model is

$$
p(x^{(1)}, \ldots, x^{(C)}, M, A) \propto \prod_{c=1}^{C} \left\{ p(A^{(c)})p(\tau^{(c)}) \left| A^{(c)} \right|^{n_c} \right\} \times \prod_{l:w_l \geq 2} \prod_{(j_1, \ldots, j_{w_l}) \in S_{M}} \left( \frac{\rho_l}{\lambda_{w_l-1}} \right) \int_{\mathcal{R}^d} \prod_{k=1}^{w_l} f(c_k)(A^{(c_k)}x_{j_k}^{(c_k)} - \mu) d\mu.
$$

In the spherical Gaussian case, again the integrals have compact explicit forms.
3-way matching
2- vs. 3-way matchings
Ongoing work

- Other transformations: affine, nonparametric ‘warping’
- Using other prior information, such as sequence numbering, local geometry, partial labelling
- Impact of maximising out geometrical transformation (in rigid-body case, with uniform priors, $\Rightarrow$ size-and-shape space)
- New MCMC samplers, effective approximations for database search
Joint work with N. Bochkina (Edinburgh)

Consistency of Bayesian estimators in generalised linear inverse problems by N. Bochkina and Peter J Green (in preparation).

Bayesian reconstructions from emission tomography data using a modified EM algorithm by Peter J Green, *IEEE Transactions on Medical Imaging*, 9, 84-93 (1990)

SPECT: single photon emission computed tomography

SPECT is a medical imaging technique for creating a 3-dimensional visualisation of the pattern of concentration of a substance of interest within the human body. As such, it images ‘function’ not ‘form’. It is complementary to other techniques such as CT and MRI/fMRI; useful for specific studies, and comparatively very cheap.

The patient is injected with/ingests/inhales a substance whose pattern of uptake within body tissue has a known relationship to physiological function. The substance has been radioactively labelled. Some of the photons subsequently emitted are detected in a gamma camera for subsequent processing and interpretation.

The reconstruction problem is to infer the pattern of concentration $x$ for the array of detected photon counts $y$; it is a statistical inverse problem, linear, but with (approximately) Poisson variation.

- modelling of SPECT was discussed,
- Bayesian methods of reconstruction were proposed, using both EM- and sampling-based computational methods,
- these methods were assessed on simulated data and real gamma camera images of both physical phantoms and actual patients.
SPECT: single photon emission computed tomography

The Poisson linear model \( y \sim \text{Poisson}(Ax) \) (component-wise, independently) is close to reality (there are some dead-time effects and other artifacts in recording).

- \( x \) represents the spatial distribution of the isotope in question, typically discretised on a square/cubic grid, \( x = \{x_s\} \).
- \( y \) the array of detected photons, also discretised \( y = \{y_t\} \) by the recording process,
- the array \( A = (a_{ts}) \) - discrete Radon transform - quantifies the emission, transmission, attenuation, decay and recording process; \( a_{ts} \) is the mean number of photons recorded at \( t \) per unit concentration at pixel/voxel \( s \).
Motivating example: pelvis data

Sinogram: raw data from a single slice, SPECT scan of pelvis
Motivating example: pelvis data
Sinogram: raw data from a single slice, SPECT scan of pelvis
In practical work, we have typically used non-Gaussian pairwise-interaction Markov random field priors:

\[
p(x) \propto \exp \left( -\beta \delta (1 + \delta) \sum_{s \sim s'} \log \cosh(\frac{x_s - x_{s'}}{\delta}) \right)
\]

This has attractive properties

- log-convex
- penalises less reconstructions \(x\) with physically-realistic abrupt boundaries (e.g. between tissue types), which are smoothed over with Gaussian priors
- bridges Gaussian (\(\delta \to \infty\)) and Laplace (\(\delta \to 0\)) Mrf’s

\(\beta\) and \(\delta\) can also be inferred.
Bayesian SPECT reconstruction

Some demonstration reconstructions... approx mle
Bayesian SPECT reconstruction

Some demonstration reconstructions... approx MAP (log cosh prior)
Bayesian SPECT reconstruction

Some demonstration reconstructions... approx MAP (log cosh prior)
There is plenty of empirical evidence of the value of taking a Bayesian approach to statistical inverse problems – but can we make any mathematical statements about the results?

Using SPECT reconstruction as an example for an asymptotic study, there are three directions in which it is very natural \( \text{in practical terms} \) to go to a limit:

- exposure time becomes large
- resolution of data becomes finer
- resolution of reconstruction becomes finer

We will concentrate on the first of these – thus we hold the dimensions of data and reconstruction fixed, but allow relative noise levels to decrease towards 0.

If the exposure time is extended by a factor \( \mathcal{T} \), the model becomes

\[
\mathcal{T} y \sim \text{Poisson}(\mathcal{T} Ax), \quad \mathcal{T} \to \infty.
\]
Inverse problem with exact data

Consider an ill-posed inverse problem: \( y = Ax \) where the columns of \( A \) are not linearly independent. Given \( y \), there is an infinite number of solutions:

\[
A^\dagger y + (I - P_A)z, \quad \forall z
\]

where \( A^\dagger \) is the Moore-Penrose inverse of \( A \), \( P_A = A^\dagger A \) and thus \((I - P_A)\) is the projection onto the kernel of \( A \).

To specify a particular solution:

\[
x^\dagger = \arg\min_{Ax = y} \|x - x_0\|_B,
\]

e.g. the solution nearest to \( x_0 \) in norm \( B \) where \( \|z\|_B = \|B^{1/2}z\| \) for some positive definite matrix \( B \). Conventionally \( x_0 = 0 \) and \( B = I \), then \( x^\dagger = A^\dagger y \).

In practice, observe \( y \) with (random?) error.
I suggest that the setting, in principle, for an inverse problem should be as follows: use all available a priori information to sequentially create models of the system, potentially an infinite number of them. For each model, solve the forward modelling problem, compare the predictions to the actual observations and use some criterion to decide if the fit is acceptable or unacceptable, given the uncertainties in the observations and, perhaps, in the physical theory being used. The unacceptable models have been falsified, and must be dropped. The collection of all the models that have not been falsified represent the solution of the inverse problem.

This concept of passing from a ‘prior collection of models’ to a ‘posterior collection of models’ will certainly be acceptable by the lovers of Bayes’ paradigm, as the collections of models can be seen as samples of a prior probability distribution and samples of a posterior distribution.
Generalised linear inverse problems

Our model for SPECT is an example of an ill-posed generalised linear inverse problem – \( p(y|x) \) depends on \( x \) only through the linear predictor \( Ax \) where \( A \) has (numerically) linearly dependent columns. We study inference for \( x \) given observed \( y \), in the limit as a noise parameter \( \sigma^2 \) (here, \( 1/T \)) goes to 0. We assume an ‘identity link function’, so that \( y \) becomes concentrated on \( Ax \) as \( \sigma^2 \to 0 \).

Because of the ill-posed/ill-conditioned character of the problem, we cannot expect consistency in inference about \( x \) based on the likelihood alone. Even as \( \sigma^2 \to 0 \), so that \( y \) converges to ‘exact data’ \( y_{\text{exact}} = Ax_{\text{true}} \), we will not be able to distinguish between \( \{ x : Ax = Ax_{\text{true}} \} \).
Generalised linear inverse problems

One of the roles of the prior in the Bayesian approach is to resolve this ambiguity (as well as generally improve reconstruction through ‘regularisation’, even without $\sigma^2 \to 0$). For the moment, we focus on the Gaussian prior $x \sim N(x_0, \gamma^2 B^{-1})$, or strictly, since we do not require $B$ to be non-singular, $p(x) \propto \exp(-1/(2\gamma^2)\|x - x_0\|_B^2)$. In the limit as $\sigma^2 \to 0$, we are interested in solutions of $Ax = y_{\text{exact}}$, where $y_{\text{exact}} = Ax_{\text{true}}$, under the influence of the prior $p(x) \propto \exp(-1/(2\gamma^2)\|x - x_0\|_B^2)$. To obtain convergence to a degenerate limit, we will need $\gamma^2 \to 0$ as well (though, as we will see, at a slower rate than $\sigma^2$).
Generalised linear inverse problems

The final ingredient we must introduce is the ‘physical’ constraint that $x$ is component-wise non-negative, that is, $x \in \mathcal{X}$, since it quantifies the isotope concentration; thus the prior is a truncated, possibly singular, Gaussian distribution.

Thus the posterior is proportional to

$$p(y|x) \times \exp\left(-\frac{1}{2\gamma^2}\|x - x_0\|_B^2\right) \text{ subject to } x \in \mathcal{X}.$$  

We will show that as $\sigma^2 \to 0$ and $\gamma^2 \to 0$ in such a way that $\nu = \gamma^2/\sigma^2 \to \infty$, the posterior converges to the point

$$x^* = \arg\min_{x \in \mathcal{X} : Ax = y_{\text{exact}}} \|x - x_0\|_B^2$$
Generalised linear inverse problems

Suppose that $A$ is a real $n \times p$ matrix, and $B$ a real symmetric non-negative definite $p \times p$ matrix, both possibly of deficient rank.

We need a rank condition to ensure that the information from the likelihood and prior together determine $x^*$ uniquely.

Assume that the $p \times 2p$ block matrix $[B: A^T A]$ has full rank $p$ (or equivalently, the rows are linearly independent). (Then for all $\nu > 0$, $B + \nu A^T A$ is nonsingular.)

It then follows that in the limit the posterior is a truncated Gaussian, with variance scaling differently in different directions. If $q$ is the rank of $A$, then asymptotically the variance of the posterior distribution (before truncation) has $q$ eigenvalues scaling like $\sigma^2$ and the remaining $(p - q)$ like the (larger) $\gamma^2$. 
Geometry

Visualisation of posterior when
\[ n = 1, \ p = 2, \ q = 1. \]
Geometry

Visualisation of posterior when 
\[
  n = 1, \ p = 2, \\
  q = 1.
\]
Geometry

Visualisation of posterior when
\[ n = 1, \ p = 3, \ q = 1. \]
In this case \( x^* \) lies internal to \( \mathcal{X} \).
Geometry

Visualisation of posterior when
\[ n = 1, \quad p = 3, \]
\[ q = 1. \]

In this case \( x^* \) lies internal to \( \mathcal{X} \).
Geometry

Visualisation of posterior when
\[ n = 1, \ p = 3, \ q = 1. \]

In this case \( x^* \) lies internal to \( \mathcal{X} \).
\( \sigma \) and \( \gamma \) getting smaller.
Geometry

Visualisation of posterior when
\[ n = 1, \quad p = 3, \]
\[ q = 1. \]

In this case \( x^* \) lies on boundary of \( \mathcal{X} \).
Geometry

Visualisation of posterior when
\[ n = 1, \ p = 3, \]
\[ q = 1. \]

In this case \( x^* \) lies on boundary of \( \mathcal{X} \).
Geometry

Visualisation of posterior when

\[ n = 1, \ p = 3, \ q = 1. \]

In this case \( x^* \) lies on boundary of \( \mathcal{X} \).

\( \sigma \) and \( \gamma \) getting smaller.
Karush-Kuhn-Tucker

We are interested in the solution to the constrained minimisation problem

\[ x^* = \arg\min_{x \in X: Ax = y_{\text{exact}}} \| x - x_0 \|_B^2 \]

By the Karush-Kuhn-Tucker theory, for this particular problem, it is necessary and sufficient to find \((x^*, \mu, \lambda) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^n\) such that

\[ B(x^* - x_0) - \mu + A^T \lambda = 0 \]

\[ x^* \geq 0 \]

\[ A x^* = y_{\text{exact}} \]

\[ \mu \geq 0 \]

for all \(s, \mu_s = 0\) or \(x_s^* = 0\)
Geometrical interpretation of KKT conditions when $n = 1$, $p = 2$, $q = 1$, in case $B = I$. 

$$Ax = y_{exact}$$
Karush-Kuhn-Tucker

Visualisation of posterior.
We have shown that the posterior distribution of $x$ in the SPECT problem, with ill-posed $A$, converges to a point mass on the point $x^* = \arg\min_{x \in \mathcal{X}} \|x - x_0\|_B^2$ as the ‘exposure time’ $1/\sigma^2 \to \infty$, providing the prior variance parameter $\gamma^2 \to 0$ as $\sigma^2 \to 0$.

The fastest rate of convergence is $\sigma^2/3 \left[-\log \sigma\right]^{1/2}$, obtained when $\gamma = \sigma^{2/3} \left[-\log \sigma\right]^{-1/3}$. If $x^*$ is on the boundary of the parameter space $x \geq 0$, there is an additional term of order $-\gamma^2 \log(c\gamma^2)$.

Approaching the limit the posterior is approximately Gaussian (with variances $\to 0$ at different rates), except that in the boundary case, it is Exponential normal to the boundary.
Sparse latent factor analysis

Joint work with S. Richardson (Imperial)

Sparse modelling of joint variation among several parallel data sets by S. Richardson and Peter J Green (in preparation).
Motivating example

In modern biology, new techniques are allowing the simultaneous probing of different stages of biological processes, so there is great interest in joint analysis of data from multiple parallel assays.


We use only data on 82 transcripts and 137 metabolites, extracting a subset where all \( p = 82 + 137 = 219 \) variables are measured under the same \( n = 18 \) conditions: 3 short time series (wildtype and a mutant unable to photosynthesise starch, under artificially prolonged diurnal and nocturnal regimes).
Motivating example: Arabidopsis data

Gibon *et al* (Genome Biology, 2006)
82 transcripts and 137 metabolites measured on 18 samples
Motivating example: Arabidopsis data

Gibon et al (Genome Biology, 2006)
82 transcripts and
137 metabolites measured on 18 samples
Main objectives

- To find a sparse representation of the dependence between sets of variables.
- Adapt latent factor models to identify co-variation, allowing for specific structure in each data set, proposing a strategy for model selection.
- ‘Open the hood’ on latent factor analysis, generally treated as a black box.
The latent factor data model

\[ X_{ji} | \cdots \sim N(\mu_j + \sum_{l=1}^{k} A_{jl} \Lambda_{li}, \Psi_j) \]

independently, for \( j = 1, 2, \ldots, p; i = 1, 2, \ldots, n \).

\( j, i \) and \( l \) index variables, samples and factors.

\( \Lambda_{li} \) are the factors and \( A_{jl} \) the factor loadings.

None of \( \mu, A, \Lambda \) or \( \Psi \) are observed/known.

We seek simultaneously latent structure, dimension reduction (\( k < (p, n) \)) and sparsity (many \( A_{jl} \) are zero).

We are interested in \( p \gg n > k \).

Existence of common patterns of loadings for multiple variables is often biologically plausible (e.g. co-expression).
The latent factor data model

Cartoon of the model

\[ X_{j|i} | \cdots \sim N(\mu_j + \sum_{l=1}^{k} A_{jl} \Lambda_{li}, \Psi_j) : \]

image plots of \( A, \Lambda, A\Lambda \) and \( X \).
Literature

Often attributed to Charles Spearman (1863–1945)
Original Bayesian formulation: Kaufman and Press, 1973

Recent explosion of interest, particularly motivated by (post-)genomic applications:

Fokoue, 2004
Carvalho, Chang, Lucas, Nevins, Wang and West, Duke TR 2005
Lucas et al, in Muller, Do, Vannucci, 2006.
Pournara and Wernisch, BMC Bioinformatics, 2007
Joint modelling: common and specific factors

When there are two blocks of variables, we can structure the factor loading matrix $A$, so there are some factors specific to each block, and some in common.

Later, we will discuss a model exploration strategy. Initially, treat this block pattern as fixed.
Prior formulation: vanilla version

Factor loadings

\[ A_{jl} | \Delta, \beta \sim N(0, \Delta_{jl}) \quad \text{where} \quad \Delta_{jl}^{-1} \sim \Gamma(\alpha_l, \beta_l) \]

independently, for \( j = 1, 2, \ldots, p; \ l = 1, 2, \ldots, k \),
where \( \alpha_l \) is fixed and \( \beta_l \) is either fixed, or \( \beta_l \sim \Gamma(e, f) \), independently, \( l = 1, 2, \ldots, k \).

Hierarchical formulation allows sharing of information, and factor-specific \((\alpha_l, \beta_l)\) adjust for differential scaling of factors.

Factors

\[ \Lambda_{li} \sim N(0, 1), \quad \text{independently, for} \ l = 1, 2, \ldots, k; \ i = 1, 2, \ldots, n. \]
Prior formulation: vanilla version

Variances

\[ \psi_j^{-1} \sim \Gamma(c, d), \text{ independently, for } j = 1, 2, \ldots, p. \]

Intercepts

\[ \mu_j \sim N(\mu_0, \tau), \text{ independently, for } j = 1, 2, \ldots, p. \]

Fixed hyperparameters

\[ \{\alpha_l\}, \{\beta_l\} \text{ or } (e, f), c, d, \mu_0, \tau \]
Prior formulation: mixture version

Factor loadings

$$A_{jl} | z, \Delta, \pi, \beta \sim \begin{cases} 0 & \text{if } z_{jl} = 0 \\ N(0, \Delta_{jl}) & \text{otherwise.} \end{cases}$$

independently, for $j = 1, 2, \ldots, p$; $l = 1, 2, \ldots, k$.

We model the mixture indicators by $z_{jl} | \pi \sim \text{Bernoulli}(\pi_l)$, independently, for $j = 1, 2, \ldots, p$; $l = 1, 2, \ldots, k$, where

$$\pi_l \sim \text{Beta}(sr, s(1 - r)),$$

independently, $l = 1, 2, \ldots, k$.

Hierarchical formulation allows sharing of information, and factor-specific ($\pi_l$) adjust for differential density of factor loadings.

Fixed hyperparameters: $\{\alpha_l\}, \{\beta_l\}$ or $(e, f), c, d, \mu_0, \tau, s, r$
MCMC computations

Blocked Gibbs sampler throughout: blocks are $\mu$, $A$, $\Lambda$, $\Psi$, $\Delta$, $\beta$, $\pi$, and all the $(A_{jl}, z_{jl})$ pairs. $A$ and $\Lambda$ use multivariate gaussian (dimension $\leq k$). Care with missing values and zeroes.

Typically 1 or $2 \times 10^5$ sweeps – about 6 minutes on laptop.
MCMC computations

Blocked Gibbs sampler throughout: blocks are $\mu$, $A$, $\Lambda$, $\Psi$, $\Delta$, $\beta$, $\pi$, and all the $(A_{jl}, z_{jl})$ pairs. $A$ and $\Lambda$ use multivariate gaussian (dimension $\leq k$). Care with missing values and zeroes. Typically 1 or $2 \times 10^5$ sweeps – about 6 minutes on laptop.
Identifiability: rotation among factors

\[ X_{ji} | \ldots \sim N(\mu_j + \sum_{l=1}^{k} A_{jl} \Lambda_{li}, \Psi_j) \]

or in matrix form

\[
X = \mu 1^T + A \Lambda + E \\
= \mu 1^T + (AR)(R^{-1} \Lambda) + E
\]

for any non-singular \( k \times k \) matrix \( R \). In particular, any rotation matrix preserves the standard normal prior form for \((R^{-1} \Lambda)\), and in the limit as \( \alpha_l, \beta_l \to \infty \) with \( \beta_l/\alpha_l \to \delta \), say, independent of \( l \), the prior distribution for \((AR)\) as well, in the vanilla case.

Even without going to this limit, if \( \alpha_l, \beta_l \) are the same for all \( l \), the priors are all invariant to a signed permutation matrix \( R \), even in the mixture case.
Identifiability: rotation among factors

\[ X = \mu \mathbf{1}^T + A\Lambda + E = \mu \mathbf{1}^T + (AR)(R^{-1}\Lambda) + E \]

This lack of identifiability is well-known and much studied, and various devices have been used to combat it.
Identifiability: rotation among factors

We monitor rotation among the factors along the MCMC run as follows: at each time point $t$ in the (thinned) MCMC sample, calculate the Procrustes rotation $U^{(t)}$ that best matches the current factor loading matrix $A^{(t)}$ to a long-run average $\bar{A}$. Plot the $k^2$ entries of $U^{(t)}$ against $t$. (In low dimensions, could replace these by generalised Euler angles).

You can assess visually whether any variation is a signed permutation between factor indices or a more general rotation.

Permutation invariance implies there will be multiple congruent modes in the posterior – no intrinsic merit in mixing across these modes if they are isolated (NB: label-switching issues).

In the case that there are permutations along the run, we can post-process the MCMC output to cancel the effect, before taking ergodic averages.
Sensitivity study: prior on $\Delta$

$A_{jl}| z, \Delta, \pi, \beta \sim N(0, \Delta_{jl})$ if $z_{jl} \neq 0$, 0 otherwise, where

$$\Delta_{jl}^{-1} \sim \Gamma(\alpha_l, \beta_l)$$

and $\alpha_l$ is fixed and $\beta_l$ is either fixed, or $\beta_l \sim \Gamma(e, f)$.

Here we will only consider $\alpha = 3$. This is the smallest (least informative) integer value for which the prior variance of $\Delta_{jl}$ is finite.

(Earlier experiments with smaller $\alpha$ showed sometimes extreme behaviour – only few high loadings preserved and even increased and the rest sent to zero, but loss of epve and some very large values for the $\Delta_{jl}$, so numerically very unstable.)

We consider $\beta = 1, 5, \sim \Gamma(1, 1)$ and $\sim \Gamma(5, 1)$. Setting - A3, $\sigma = 2$. Mixture prior, $s = 50, r = 0.1$. 
Sensitivity study: prior on $\Delta$

We find sensitivity of posterior mean loadings $E(A_{jl}|X)$.

$\beta = 1$ vs. $\beta = 5$

left vs. right

Recall $E(\Delta_{jl}) = \beta_l/(\alpha_l - 1)$
Sensitivity study: prior on $\Delta$

... that is alleviated by using hierarchical setting.

$\beta \sim \Gamma(1, 1)$ vs. $\beta \sim \Gamma(5, 1)$

left vs. right
Model choice strategy

- How many factors?
- How many factors specific to each block of variables, and how many common ones?

Our current recommendation is to use a stepwise, backwards-selection, non-automated approach, examining
  - sparsity and goodness of fit summaries
  - patterns of non-zero loadings $p_{jl}$
  - correlations between factors

We demonstrate this on a new simulated dataset (A7z) and on the Arabidopsis data.
A7z simulated data

In this set-up, there are 100 variables in each block, with 2 and 3 factors specific to the blocks, respectively, and 2 common factors (the ‘2+3+2’ model). The noise has $\sigma = 2$. 
Sparse latent factor analysis

Model choice

Strategy applied to A7z simulated data

11 factors: 5+6+0. We begin with plenty of specific factors and no common ones. One factor in each block has no $\text{pp}_{jl} > 0.5$, so drop to . . .

9 factors: 4+5+0. All factors have many $\text{pp}_{jl} > 0.5$, but there is strong correlation between profiles of 2 factors, 1 in each block: suggests replacing them by a common factor.
Strategy applied to A7z simulated data

8 factors: 3+4+1. Again, all factors have many $p_{jl} > 0.5$, but there is strong correlation between profiles of 2 factors, 1 in each block: again replace by a common factor...

7 factors: 2+3+2. No pair of factors has correlation $> 0.52$, so we stop. This was the model used to generate the data.
Strategy applied to A7z simulated data

Final model: 7 factors, 2&3 specific + 2 common, as generated

Posterior mean loadings (y-axis) vs. true loadings

epve = 73.0%, epnz = 34.0%
Strategy applied to A7z simulated data

True and estimated factor profiles
Return to motivating example: Arabidopsis data

Gibon et al (Genome Biology, 2006)
82 transcripts and
137 metabolites measured on 18 samples
Sparse latent factor analysis
Arabidopsis transcripts and metabolites

Strategy applied to Arabidopsis data

12 factors: 6+6+0. Three factors had no \( pp_{jl} > 0.1 \) even, so drop to . . .

9 factors = 6+3+0. All have plenty of \( pp_{jl} > 0.5 \); strongest correlation (0.936) between factors 1 & 7, so replace by a common factor . . .
Strategy applied to Arabidopsis data

8 factors = 5+2+1. Factors 1 and 6 should be combined; also factor 7 is very sparse, so it seems no specific factors are needed for first block.

6 factors = 4+0+2. High correlation between factor 1 (specific) and 6 (common) so eliminate a specific factor.
Strategy applied to Arabidopsis data

5 factors = 3+0+2. High correlation between factor 3 (specific) and 5 (common) so eliminate another specific factor...

4 factors = 2+0+2. No factors are sparse, no pairs of factors highly correlated.
Strategy applied to Arabidopsis data

Fit of final model:
- 4 factors,
- 2 common to transcripts and metabolites,
- 2 specific to metabolites

Suggestive of metabolites regulating transcripts, rather than v.v.?

$epve = 59.4\%$, $epnz = 36.1\%$
Strategy applied to Arabidopsis data

$pp_{jl}$ for final model, thresholded at 0.5 – 4 MCMC runs, different starts
Strategy applied to Arabidopsis data

\[ p_{jl} \] for final model, thresholded at 0.8 – 4 MCMC runs, different starts
Profiles 1 and 2 essentially capture the distinction between the Wildtype and the *pgm* starchless mutant, while 3 and 4 are distinguished by the light condition (day versus prolonged night).
Strategy applied to Arabidopsis data – interpretation of common factors

Both involve amino acid and carbohydrate metabolism. Profile 3: increase of a group of amino acids (isoleucine, leucine, valine, tyrosine, phenylalanine, tryptophan) during the prolonged night together with transcripts (NR, GS, Fd-GOGAT, Asp AT and Ala AT) involved in amino acid metabolism. Profile 4: increase of several carbohydrates (e.g. UDP glucose, sucrose), glycolysis (Pyruvate) and organic acids (Fumarate, malate) during the day together with associated transcripts (fructokinase, glucokinase, pyruvate kinase, NAD-GAPDH, PK,G6PDH, PFP) for both WT and the pgm mutant. Future?: interpret metabolites/transcripts of unknown activity.
Discussion

- Factor analysis seems promising as a route to uncovering sparse patterns of dependence between $\geq 2$ groups of variables.
- We have identified prior settings that we believe makes the model flexible without compromising performance.
- Our model choice strategy probably needs further exploration and refinement, and possible automation.
- There is a need for a model choice criterion balancing sparsity and fit.
- It is necessary to pay attention to scaling of variables.
- We encountered no serious MCMC difficulties on problems of this scale, although of course permutations and sign changes need to be accommodated; we are optimistic about scalability.